Annotating Character Relationships in Literary Texts

Philip Massey and Patrick Xia
School of Computer Science
Carnegie Mellon University
{pmassey,paxia}@cmu.edu

David Bamman
School of Information
University of California, Berkeley
dbamman@berkeley.edu

Noah A. Smith
Computer Science & Engineering
University of Washington
nasmith@cs.washington.edu

1 Overview

We present a dataset of manually annotated relationships between characters in literary texts, in order to support the training and evaluation of automatic methods for relation type prediction in this domain (Makazhanov et al., 2014; Kokkinakis, 2013) and the broader computational analysis of literary character (Elson et al., 2010; Bamman et al., 2014; Vala et al., 2015; Flekova and Gurevych, 2015). In this work, we solicit annotations from workers on Amazon Mechanical Turk on four dimensions of interest: for a given pair of characters, we collect judgments as to the **coarse-grained category** (professional, social, familial), **fine-grained category** (friend, lover, parent, rival, employer), and **affinity** (positive, negative, neutral) that describes their primary relationship in a text. We do not assume that this relationship is static; we also collect judgments as to whether it **changes** at any point in the course of the text.

The annotations describe character dyads in 109 texts ranging from Homer's *Iliad* to Joyce's *Ulysses* (see section 3 for a full list). Rather than relying on annotators' expertise in these texts, we frame the annotation problem as one of estimating the relationship as depicted in a third-party summary (SparkNotes); this allows annotators to provide judgments on the relationship between pairs of characters by only reading a summary of a book, and not the book itself. While this approach naturally loses the nuance of a truly expert opinion, it allows us to broadly characterize a large number literary dyads, assess the feasibility of this annotation strategy, and provide a foundation on which other work can build. All data is openly available at http://github.com/dbamman/characterRelations, and we encourage contributions and corrections.

2 Data Collection

Our primary corpus selection criterion was that a text is both available on Project Gutenberg (to enable computational analysis of an open-access text) and is the subject of a study guide on www.sparknotes.com (to enable annotation by non-experts). SparkNotes provides a detailed summary of the plot and major characters in texts, often a structured format (e.g., a section denoted "Character List"); from this summary we extract all characters and use them to populate the following questionnaire.

Fictional Character Relationship Analysis

In this task, you'll be identifying the relationship type that exists between two characters in *The Good Soldier* by Ford Madox Ford, using a description and summary of that work from SparkNotes. (For example, if reading *To Kill a Mockingbird*, you'd mark that Atticus Finch is the father of Scout Finch.)

We expect this task to take approximately 20 minutes. Please read the "Character List," "Plot Overview," and "Character Analysis / Analysis of Major Characters" pages here: `http://www.sparknotes.com/lit/goodsoldier/`.

After reading these pages, list all of the relationships that you can identify between the characters described there. To complete a relationship, find the two characters in the dropdown menus below with First Character and Second Character. Then, answer the following questions using these guidelines:

Affinity:

How do the two characters feel toward each other? For example, if they are friendly, select "Positive". If it is unclear how they feel toward each other or if they do not have strong opinions about each other, select "Neutral". If they are enemies or rivals, or hate each other, select "Negative".

Category:

How are the two characters related? If they are friends, select "Social". If they share a relationship because they work with each other, select "Professional". If they are family, then select "Familial".

Kind:

Specifically, how is the First Character related to the Second Character? For example, if the First Character is the husband of the Second Character, then select "husband". Please keep the Category and the Kind consistent:

If you selected "Social", then the Kind must be one of these:

- friend
- enemy
- acquaintance
- lovers
- unrequited love interest (X is in love with Y, but Y is not in love with X)
- rivals

If you selected "Professional", then the Kind must be one of these:

- employer
- employee
- colleague
- servant
- master
- student
- teacher
- client

- person offering service to client (e.g., lawyer)

If you selected "Familial", then the Kind must be one of these:

- husband/wife
- brother/sister
- cousin
- uncle/aunt
- niece/nephew
- child
- parent
- grandchild
- grandparent
- orphan
- foster parent
- step-child
- step-parent
- in-law relation (e.g., mother-in-law; specify in detail)
- half relation (e.g., half-sister; specify in detail)

Change:

Does the relationship between the two characters significantly change at some point in the book? For example, does a "positive" relationship become "negative", or do "lovers" become "husband/wife" or "friends" become "enemies"? If so, select "Yes" here and describe the change in the Detail section below. If the relationship remains the same throughout (for example, "brother/sister"), then select "No".

Detail:

If the relationship type you feel holds between two characters was not provided, or you want to provide additional information, enter it here.

If two characters hold multiple relations to each other (such as "Professional: colleague" and "Social: lovers"), enter those multiple relations in different rows below. Please identify a total of 10 relations between characters in the rows below.

(Users then select a pair of characters from the following list and label the relations specified above):

- John Dowell
- Florence Hurlbird Dowell
- Leonora Powys Ashburnham
- Captain Edward Ashburnham
- Nancy Rufford
- Jimmy
- Uncle John Hurlbird
- Maisie Maidan
- Rodney Bayham
- Mrs. Basil
- La Dolciquita

- The Misses Hurlbird
- Selmes
- Major Rufford

We present this questionnaire to workers on Amazon Mechanical Turk, soliciting two independent judgments for each of the 109 literary texts. Since we are soliciting judgments regarding any 10 character pairs in the text (and not a fixed set of such dyads), many of the character pairs from different annotators for the same work do not overlap. We collect a total of 2,170 annotations; among these, 392 character dyads have annotations by two different annotators, from which we can calculate agreement statistics. Table 1 lists the inter-annotator agreement rate (and Fleiss' κ, correcting for chance) for each of the four annotation classes. The agreement rates for both the coarse- and fine-grained categories are both high, even when correcting for chance ($\kappa = 0.812$ and 0.744, respectively). 18.1% of character pairs are judged to exhibit some change over the course of the text. While annotators display high agreement on this (75.7%), their agreement is in fact quite low when correcting for chance ($\kappa = 0.208$). Judging whether a dyad's relationship is *positive, negative* or *neutral* also proves to be quite difficult, with low agreement rates across annotators.

Figure 1 shows the distribution of affinity and coarse-grained category annotations, while figure 2 shows the distribution of fine-grained category annotations.

Label type	IAA	Fleiss' κ	n
Affinity	0.627	0.364	391
Category	0.879	0.812	389
Kind	0.765	0.744	392
Change	0.757	0.208	371

Table 1: Inter-annotator agreement (IAA) rates, along with chance-corrected Fleiss' κ for the four annotation tasks. n denotes the sample size of directed character pairs with two annotations; sample size exhibits question-level variability due to incomplete responses.
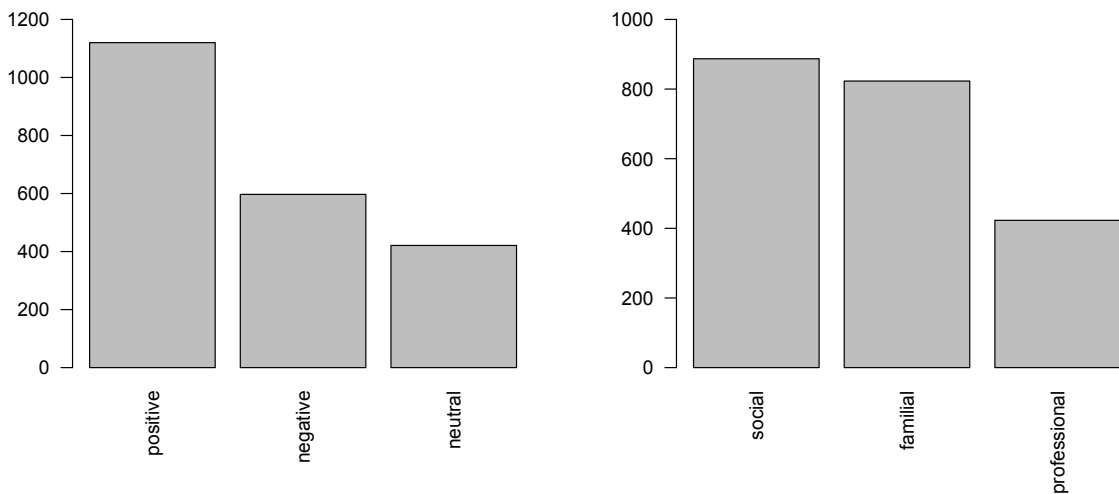

Figure 1: Distribution of affinity annotations (left) and coarse-grained category annotations (right).
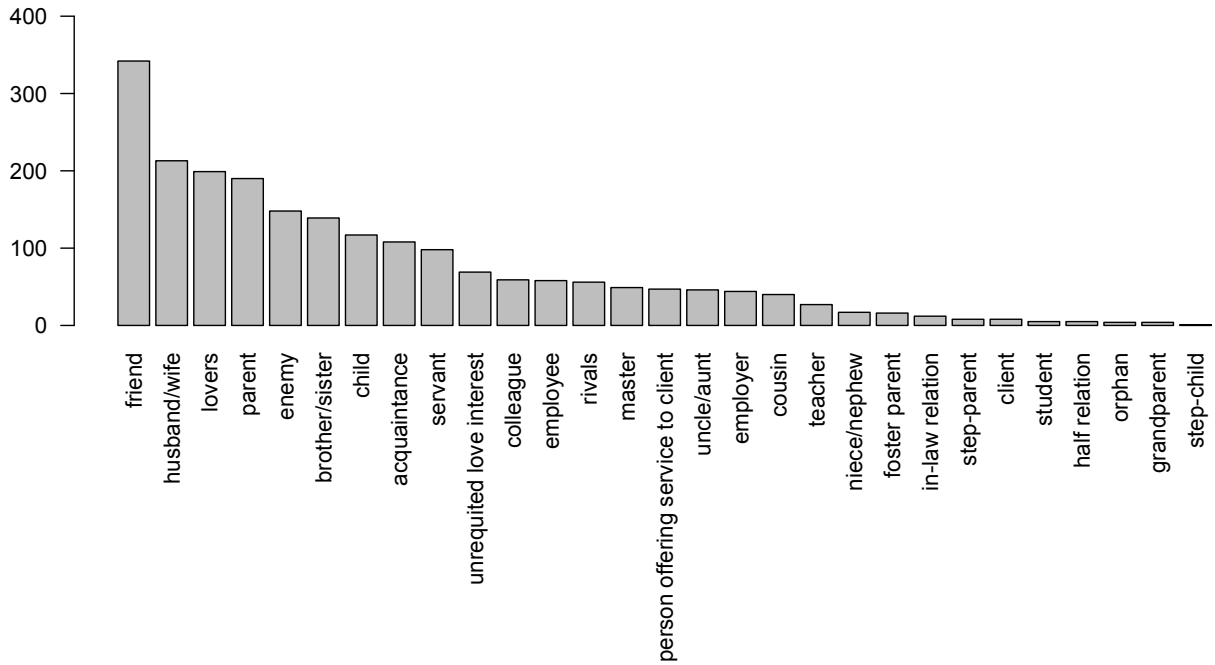
Figure 2: Distribution of fine-grained category annotations.

3 Texts

- Alexandre Dumas. The Count of Monte Cristo, The Three Musketeers
- Aristophanes. Lysistrata
- Bram Stoker. Dracula
- Charles Dickens. A Tale of Two Cities, Bleak House, Great Expectations, Hard Times, Oliver Twist
- Charlotte Perkins Gilman. Herland
- Christopher Marlowe. The Jew of Malta
- E. M. Forster. Howards End
- Edith Wharton. Ethan Frome, The House of Mirth
- Edmond Rostand. Cyrano de Bergerac
- F. Scott Fitzgerald. This Side of Paradise
- Frances Hodgson Burnett. The Secret Garden
- Franz Kafka. The Trial
- Frederick Douglass. Narrative of the Life of Frederick Douglass
- George Bernard Shaw. Major Barbara, Pygmalion
- George Eliot. Adam Bede, Middlemarch, Silas Marner
- Gustave Flaubert. Madame Bovary
- Harriet Beecher Stowe. Uncle Tom's Cabin
- Henrik Ibsen. A Doll's House, Ghosts, Hedda Gabler
- Henry James. The American, The Portrait of a Lady, The Turn of the Screw

- Herman Melville. Typee
- Hermann Hesse. Siddhartha
- Homer. The Iliad
- Jack London. White Fang
- James Fenimore Cooper. The Last of the Mohicans
- James Joyce. A Portrait of the Artist as a Young Man, Dubliners, Ulysses
- Jane Austen. Emma, Mansfield Park, Northanger Abbey, Persuasion, Pride and Prejudice, Sense and Sensibility
- John Milton. Paradise Lost
- Joseph Conrad. Heart of Darkness, Lord Jim
- L. M. Montgomery. Anne of Green Gables
- Leo Tolstoy. Anna Karenina, War and Peace
- Louisa May Alcott. Little Women
- Marcel Proust. Swann's Way
- Mark Twain. The Adventures of Tom Sawyer
- Miguel de Cervantes. Don Quixote
- Nathaniel Hawthorne. The House of the Seven Gables, The Scarlet Letter
- Oscar Wilde. An Ideal Husband, The Picture of Dorian Gray
- Plato. The Republic
- Robert Louis Stevenson. Dr. Jekyll and Mr. Hyde, Kidnapped, Treasure Island
- Sinclair Lewis. Babbitt, Main Street
- Stephen Crane. Maggie: A Girl of the Streets
- Theodore Dreiser. Sister Carrie
- Thomas Hardy. Far from the Madding Crowd, Jude the Obscure, The Mayor of Casterbridge, The Return of the Native
- Thomas Kyd. Spanish Tragedy
- Upton Sinclair. The Jungle
- Virgil. The Aeneid
- Voltaire. Candide
- Willa Cather. O Pioneers!
- William Shakespeare. A Midsummer Night's Dream, Antony and Cleopatra, As You Like It, Hamlet, Henry IV Part 1, Henry IV Part 2, Henry V, Henry VI Part 1, Henry VI Part 2, Henry VI Part 3, Henry VIII, Julius Caesar, King John, King Lear, Love's Labours Lost, Macbeth, Measure for Measure, Much Ado About Nothing, Othello, Richard II, Romeo and Juliet, The Comedy of Errors, The Merchant of Venice, The Merry Wives of Windsor, The Taming of the Shrew, The Tempest, The Two Gentlemen of Verona, The Winter's Tale, Timon of Athens, Titus Andronicus, Troilus and Cressida, Twelfth Night

4 Data

All annotations are freely available at http://github.com/dbamman/characterRelations. Rather than reconciling disagreements between annotators or filtering out incomplete responses, we are making all collected data available; each annotations is paired with an (anonymized) identifier of the annotator who provided it.

5 Acknowledgments

6 References

David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 138–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Lucie Flekova and Iryna Gurevych. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Dimitrios Kokkinakis. Annotation of interpersonal relations in Swedish prose fiction. In *Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, 2013.

Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. Extracting family relationship networks from novels. *CoRR*, abs/1405.0603, 2014.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal, September 2015. Association for Computational Linguistics.